Prediction of toxicity of Ionic Liquids based on GC-COSMO method

Peng, Daili; Picchioni, Francesco

Published in:
Journal of hazardous materials

DOI:
10.1016/j.jhazmat.2020.122964

Publication date:
2020

Journal of Hazardous Materials 398 (2020) 122964

Contents lists available at ScienceDirect

Journal of Hazardous Materials

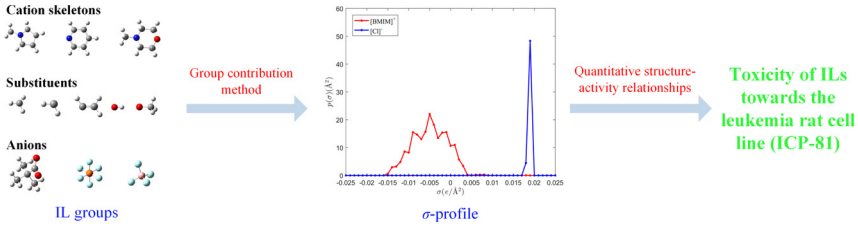
journal homepage: www.elsevier.com/locate/jhazmat

Prediction of toxicity of Ionic Liquids based on GC-COSMO method

Daili Peng, Francesco Picchioni*

University of Groningen, Faculty of Science and Engineering, Product Technology – Engineering and Technology Institute Groningen, Nijenborgh 4, 9747 AG Groningen, the Netherlands

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: D. Aga

Keywords:
QSAR
COSMO-SAC
group contribution
toxicity
ILs
CAILD

ABSTRACT

In order to evaluate the toxicity of several different ionic liquids (ILs) towards the leukemia rat cell line (ICP-81), an efficient and reliable quantitative structure-activity relationships (QSAR) model is developed based on descriptors from COSMO-SAC (conductor-like screening model for segment activity coefficient) model. The distribution of screen charge density (σ-profile) of 127 ILs is calculated by GC-COSMO (group contribution based COSMO) method. Two segmentation methods toward σ-profile are used to find out the appropriate descriptors for the QSAR model. The optimal subset of descriptors is obtained by enhanced replacement method (ERM). A multiple linear regression (MLR) and multilayer perceptron technique (MLP) are used to build the linear and nonlinear models, respectively, and the applicability domain of the models is assessed by the Williams plot. It turns out that the nonlinear model based the second segmentation method (MLP-2) is the best QSAR model with an $R^2 = 0.975$, $MSE = 0.019$ for the training set and $R^2 = 0.938$, $MSE = 0.037$ for the test set. The reliability and robustness of the presented QSAR models are confirmed by Leave-One-Out (LOO) cross and external validations.

1. Introduction

With their unique properties, such as negligible vapor pressure, high thermal, chemical stability and wide liquid-phase range, ionic liquids (ILs) have been researched for a diverse range of technologies and applications, including gas capture and separation (Bates et al., 2002; Chen et al., 2015), extraction (Lyu et al., 2014; Song et al., 2016; Wlazło et al., 2017; Zhou et al., 2012), organic synthesis (Eshetu et al., 2016; Sanchez Zayas et al., 2016), etc. Moreover, because of their low volatility, atmospheric pollution is unlikely; thus ILs are widely considered as “green” solvents compared to traditional volatile organic compounds (VOCs). However, it is now realized that ILs have hazard potentials for the human being and the environment (Ventura et al., 2013). Due to their significant solubility in water, the possible industrial discharge of wastewater containing ILs into the environment may have detrimental toxicological consequences for aquatic organisms (Singh et al., 2014). On the other hand, the properties of ILs, such as thermal stability and non-volatility, might also pose environmental threats because of slow degradation characteristic (Cao et al., 2018). In order to find environmentally friendly ILs for different using purposes, evaluation of their toxicity has become very important.

In principle, there are approximately 10^{18} anion-cation

* Corresponding author.
E-mail address: f.picchioni@rug.nl (F. Picchioni).

https://doi.org/10.1016/j.jhazmat.2020.122964
Received 22 December 2019; Received in revised form 22 April 2020; Accepted 14 May 2020
Available online 25 May 2020

combinations that can be synthesized (Huang et al., 2013). To avoid the time and labor intensive experiment, many QSAR (quantitative structure-activity)/ QSPR (quantitative structure-property) prediction models have been built to predict the thermophysical properties of ILs, such as melting points (Lazzús, 2012), surface tensions (Gharagheizi et al., 2012), viscosities (Lazzús and Pulgar-Villarroel, 2015), glass transition temperature (Mirkhani et al., 2012), decomposition temperature (Yan et al., 2012). As to the models for toxicity prediction, they can be divided into two categories according to the descriptor used to build the model. Group Contribution (GC) based methods directly use the frequency of IL groups to predict the toxicity (Hossain et al., 2011; Luis et al., 2010, 2007). The main advantage of GC is its simplicity and capability to give a reasonable accuracy if all the necessary group increments are obtained from the experimental data (Chen et al., 2013). Moreover, GC-based methods can be directly integrated into the computer-aided ionic liquid design (CAILD) framework.

Another category of models is based on the descriptors that have certain connection to the characteristic of ILs rather than the frequency of groups, e.g. the topological index (García-Lorenzo et al., 2008; Yan et al., 2015), free energy relationship (Cho et al., 2013) and the distribution of screen charge density (Ghanem et al., 2017; Torrecilla et al., 2010). The distribution of screen charge density distribution is also referred as the σ-profile and can be achieved by COSMO computation. The σ-profile is considered as a characteristic property of the molecule; it can be used to predict the possible electrostatic, hydrogen-bonding, and dispersion interactions of the compound. Different descriptors based on σ-profile of COSMO-RS model have been successfully used to build the QSAR models for estimating the toxicity of ILs. Ghanem et al. (2017) divided the σ-profile of cation and anion into four regions separately. The area under each region is regarded as the descriptor that is used to build the QSAR model for predicting the ecotoxicity of 110 ILs towards bioluminescent bacterium *Vibrio fischeri*. The squared correlation coefficient (R^2) and mean square error (MSE) of the nonlinear model using MLP model are 0.961 and 0.157, respectively. Torrecilla et al. (2010) treated the charge distribution area ($S_{\sigma\text{-profile}}$) below the σ-profile as the descriptor. Because the σ-profile of COSMO-RS model is from -0.03 to 0.03 with a step size of 0.001, there are 61 $S_{\sigma\text{-profile}}$ descriptors for each cation and anion. After the regression model selection (RMS) analysis, 10 out of 102 descriptors are chosen to build the QSAR model for predicting the toxicity of 105 ILs towards leukemia rat cell line (ICP-81) ($R^2 > 0.996$ for the final MLP model). Although these methods can achieve satisfying results, they still have room for improvement. First, the quantum mechanical calculations for generating the σ-profile are very time-consuming and computationally expensive (Mullins et al., 2006). Secondly, in these methods IL is treated as an ion pair rather than individual functional groups, which makes them hard to be integrated into the CAILD framework.

In order to take the advantage of using σ-profile as the descriptor and provide a fast and reliable prediction method for the toxicity of ILs towards ICP-81, which can be used for CAILD, GC-COSMO (Group contribution based COSMO) is used in this work to predict the σ-profile of ILs for COSMO-SAC model. Two segmentation methods for σ-profile from literature are compared in order to find out the suitable descriptors. The optimal set of descriptors are selected by ERM (Enhanced replacement method) and used to build the linear and nonlinear QSAR models using MLR (Multi-Linear Regression) and MLP (Multi-Layer Perceptron technique), respectively. The performances of the obtained QSAR models are then investigated and compared with previous studies.

2. Methodology

The strategy of the presented method is illustrated in Fig. 1. Firstly, a database covering the information of σ-profile for different IL groups is obtained from our previous work (Peng et al., 2017). Then, the σ-profile of the ILs are calculated based on the GC-COSMO method. After that, the descriptors are calculated by two segmentation methods for σ-profile and the optimal set of descriptors are derived from ERM algorithm. Finally, MLR and MLP are used to build the linear and nonlinear QSPR models for each segmentation method.

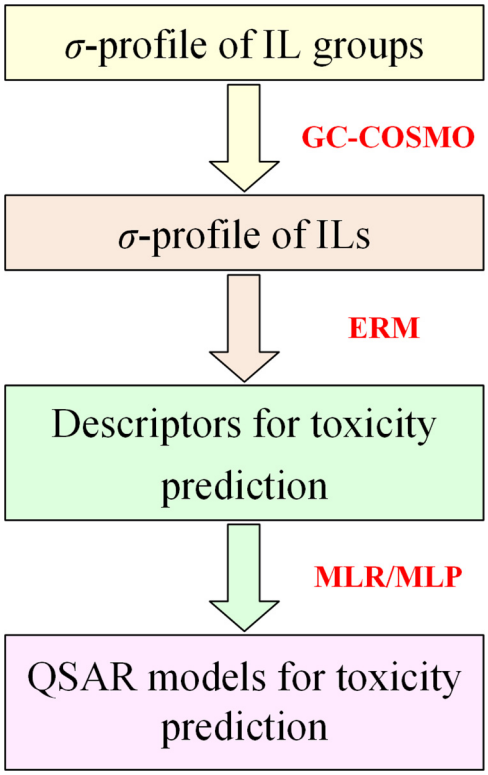


Fig. 1. Framework of the proposed method for the prediction of toxicity of ILs.

2.1. Dataset

In order to compare with the recent research for the prediction of toxicity of ILs, the same training and test set used by Cao et al. (2018) are employed in this work. The toxicity data of the chosen ILs is from the widely acknowledged ILs database (The UFT/ Merck Ionic Liquids Biological Effects Database, 2020; Zhang et al., 2006). It is worth noting that 7 ILs are excluded from the original dataset because their group information is temporarily not included in the GC-COMSO database. In addition, 15 new ILs from different databases (Ranke et al., 2004; The UFT/ Merck Ionic Liquids Biological Effects Database, 2020; Torrecilla et al., 2009) are added to the original dataset as an external validation set to further evaluate the predictive ability of the developed models. This choice is justified on one hand by the use of a common dataset for the model development (see above) and on the other one by randomly selecting 15 IL as the validation set. Therefore, 127 ILs are included in the dataset with 93 ILs as the training set, 19 ILs as the test set and 15 ILs as the validation set. The name and the experimental log$EC50$ value of ILs are listed in Table 1.

2.2. GC-COSMO method

In the GC-COSMO method (Peng et al., 2017), ILs are decomposed into three parts (Fig. 2): cation skeleton, substitutes in the cation skeleton and anion. As seen in Fig. 3a, for every group, the σ-profile is defined as a vector of 51 elements from -0.025 to 0.025 with a step size of 0.001. The σ-profile of anion can be directly acquired by GC-COSMO method since it is regarded as one group. The σ-profile of cation is defined as the accumulation of σ-profile from cation skeleton and its substitutes (Fig. 3b):

Table 1
The experimental versus calculated log EC50 values using different models.

No.	Cations	Anions	Exp.	MLR-1	MLP-1	MLR-2	MLP-2
1	1-(3-methoxypropyl)-1-methylpiperidinium	chloride	4.40	3.93	4.13	4.08	4.56
2	1-(3-methoxypropyl)-1-methylpiperidinium	bis(trifluoromethylsulfonyl)amide	3.27	3.18	3.26	3.49	3.36
3	1-benzyl-3-methylimidazolium	tetrafluoroborate	2.97	3.05	2.93	3.32	2.95
4	1-butyl-1-methylpiperidinium	bromide	4.03	3.63	3.76	3.91	4.22
5	1-butyl-1-methylpiperidinium	bis(trifluoromethylsulfonyl)amide	3.41	2.93	3.16	3.31	3.47
6	1-butyl-3-methylimidazolium	2-(2-methoxyethoxy)ethylsulfate	3.15	3.05	3.10	2.92	3.14
7	1-butyl-3-methylimidazolium	bromide	3.43	3.50	3.51	3.39	3.33
8	1-butyl-3-methylimidazolium	chloride	3.55	3.55	3.73	3.39	3.33
9	1-butyl-3-methylimidazolium	iodide	3.48	3.43	3.21	3.39	3.33
10	1-butyl-3-methylimidazolium	bis(trifluoromethylsulfonyl)amide	2.68	2.80	2.77	2.80	2.81
11	1-butyl-3-methylpyridinium	tetrafluoroborate	3.30	3.17	3.17	2.86	3.12
12	1-butylpyridinium	tetrafluoroborate	3.16	3.32	3.49	3.25	3.19
13	1-butylpyridinium	bromide	3.90	3.55	3.70	3.66	3.69
14	1-butylpyridinium	chloride	3.77	3.60	3.83	3.66	3.69
15	1-butylpyridinium	methylsulfate	3.92	3.32	3.50	3.73	3.70
16	1-butylpyridinium	trifluoromethanesulfonate	3.66	3.16	3.58	3.53	3.78
17	1-ethyl-3-methylimidazolium	acetate	4.23	4.12	4.15	3.91	4.00
18	1-ethyl-3-methylimidazolium	tetrafluoroborate	3.44	3.86	3.78	3.64	3.42
19	1-ethyl-3-methylimidazolium	methanesulfonate	3.97	3.91	3.82	4.10	4.08
20	1-ethyl-3-methylimidazolium	trifluoroacetate	4.00	3.81	4.03	3.98	4.08
21	1-ethyl-3-methylimidazolium	trifluoromethanesulfonate	4.09	3.69	3.80	3.93	4.15
22	1-heptyl-3-methylimidazolium	chloride	2.53	2.67	2.35	2.50	2.51
23	1-hexadecyl-3-methylimidazolium	chloride	−0.24	−0.62	−0.19	−0.17	−0.37
24	1-hexyl-1-methylpyrrolidinium	chloride	2.93	3.30	3.18	2.97	3.00
25	1-hexyl-1-methylpyrrolidinium	bis(trifluoromethylsulfonyl)amide	2.56	2.55	2.66	2.37	2.41
26	3-hexyl-1,2-dimethylimidazolium	tetrafluoroborate	1.90	2.67	1.87	2.13	1.99
27	1-hexyl-3-methylpyridinium	chloride	2.40	2.69	2.69	2.68	2.63
28	1-hexyl-4-methylpyridinium	tetrafluoroborate	2.17	2.40	2.21	2.22	2.19
29	1-hexyl-4-methylpyridinium	chloride	2.67	2.68	2.64	2.63	2.60
30	1-hexylpyridinium	chloride	2.80	3.06	2.88	2.97	2.77
31	1-hexylpyridinium	trifluoromethanesulfonate	2.54	2.62	2.65	2.84	2.54
32	3-methyl-1-nonylimidazolium	chloride	1.40	2.01	1.44	1.91	1.60
33	1-methyl-1-octylpyrrolidinium	chloride	2.59	2.55	2.46	2.38	2.31
34	3-methyl-1-octylimidazolium	tetrafluoroborate	1.59	2.09	1.64	1.79	1.85
35	3-methyl-1-octylimidazolium	chloride	2.00	2.37	1.84	2.21	2.08
36	3-methyl-1-octylimidazolium	bis(trifluoromethylsulfonyl)amide	1.64	1.62	1.48	1.61	1.56
37	1-methyl-3-pentylimidazolium	chloride	3.16	3.26	3.32	3.10	3.11
38	3-methyl-1-propylimidazolium	tetrafluoroborate	3.45	3.57	3.51	3.29	3.40
39	1-octyl-4-methylpyridinium	tetrafluoroborate	1.49	1.65	1.50	1.62	1.34
40	1-butyl-4-methylpyridinium	chloride	3.32	3.43	3.29	3.23	3.31
41	1-(2-ethoxyethyl)-1-methylpiperidinium	bis(trifluoromethylsulfonyl)amide	3.34	3.20	3.23	3.37	3.28
42	1-(2-ethoxyethyl)pyridinium	bis(trifluoromethylsulfonyl)amide	3.26	3.09	3.29	3.19	3.31
43	1-(2-hydroxyethyl)-1-methylpiperidinium	iodide	4.58	4.43	4.37	4.43	4.68
44	1-(2-hydroxyethyl)-1-methylpiperidinium	bis(trifluoromethylsulfonyl)amide	3.65	3.80	3.69	3.83	3.64
45	1-(2-hydroxyethyl)pyridinium	iodide	4.16	4.28	4.31	4.25	4.20
46	1-(2-methoxyethyl)-1-methylpiperidinium	bis(trifluoromethylsulfonyl)amide	3.25	3.40	3.38	3.51	3.33
47	1-(3-hydroxypropyl)-1-methylpyrrolidinium	bis(trifluoromethylsulfonyl)amide	3.60	3.85	3.67	3.67	3.62
48	1-(3-hydroxypropyl)-1-methylpyrrolidinium	bis(trifluoromethylsulfonyl)amide	3.62	3.85	3.67	3.67	3.62
49	1-(3-methoxypropyl)pyridinium	bis(trifluoromethylsulfonyl)amide	3.38	3.08	3.28	3.30	3.42
50	1-(cyanomethyl)-1-methylpiperidinium	bis(trifluoromethylsulfonyl)amide	3.95	3.80	3.59	3.92	3.93
51	1-(ethoxymethyl)-1-methylpiperidinium	chloride	4.24	4.20	4.14	3.95	4.09
52	1-(ethoxymethyl)-1-methylpiperidinium	bis(trifluoromethylsulfonyl)amide	3.41	3.45	3.40	3.35	3.24
53	1-(ethoxymethyl)pyridinium	chloride	3.32	4.09	3.26	3.83	3.71
54	1-pentylpyridinium	bromide	3.15	3.28	3.27	3.32	3.27
55	1-pentylpyridinium	bis(trifluoromethylsulfonyl)amide	2.85	2.58	2.55	2.72	2.83
56	1-propylpyridinium	bis(trifluoromethylsulfonyl)amide	3.20	3.12	3.41	3.41	3.21
57	4-(2-ethoxyethyl)-4-methylmorpholinium	bis(trifluoromethylsulfonyl)amide	3.69	3.36	3.52	3.55	3.60
58	4-(2-methoxyethyl)-4-methylmorpholinium	bis(trifluoromethylsulfonyl)amide	3.81	3.55	3.70	3.69	3.69
59	4-(3-hydroxypropyl)-4-methylmorpholinium	bis(trifluoromethylsulfonyl)amide	3.53	3.72	3.93	3.78	3.80
60	4-(3-methoxypropyl)-4-methylmorpholinium	bis(trifluoromethylsulfonyl)amide	3.77	3.26	3.71	3.34	3.59
61	4-butyl-4-methylmorpholinium	bis(trifluoromethylsulfonyl)amide	3.43	2.99	3.20	3.15	3.34
62	4-(ethoxymethyl)-4-methylmorpholinium	bis(trifluoromethylsulfonyl)amide	3.36	3.63	3.58	3.62	3.62
63	4-ethyl-4-methylmorpholinium	toluene-4-sulfonate	3.81	3.89	3.81	4.04	3.86
64	benzyltetradecyldimethylammonium	chloride	0.16	−0.41	0.23	−0.20	0.12
65	1-(2-ethoxyethyl)-1-methylpyrrolidinium	bis(trifluoromethylsulfonyl)amide	3.20	3.44	3.26	3.24	3.33
66	1-(2-hydroxyethyl)-3-methylimidazolium	bis(trifluoromethylsulfonyl)amide	3.76	3.73	3.68	3.53	3.54
67	1-(2-methoxyethyl)-1-methylpyrrolidinium	bis(trifluoromethylsulfonyl)amide	3.30	3.64	3.41	3.39	3.44
68	1-(2-methoxyethyl)-3-methylimidazolium	bis(trifluoromethylsulfonyl)amide	3.25	3.30	3.24	3.21	3.34
69	1-butyl-1-methylpyrrolidinium	bromide	3.77	3.86	3.83	3.64	3.75
70	1-butyl-1-methylpyrrolidinium	dicyanamide	4.23	3.62	3.92	3.75	3.84
71	1-butyl-1-methylpyrrolidinium	bis(trifluoromethylsulfonyl)amide	3.01	3.15	3.12	3.04	2.87
72	1-butyl-1-methylpyrrolidinium	trifluorotris(pentafluoroethyl)phosphate	2.41	2.85	2.52	2.46	2.35
73	1-butyl-3-ethylimidazolium	trifluoroacetate	3.31	3.01	3.26	3.06	3.32
74	1-butyl-3-ethylimidazolium	trifluoromethanesulfonate	3.43	2.89	3.32	3.01	3.39

(*continued on next page*)

Table 1 (*continued*)

No.	Cations	Anions	Exp.	MLR-1	MLP-1	MLR-2	MLP-2
75	1-butyl-3-methylimidazolium	dicyanamide	3.15	3.27	3.27	3.51	3.43
76	1-butyl-3-methylimidazolium	hydrogensulfate	3.29	3.28	3.29	3.55	3.23
77	1-butyl-3-methylimidazolium	methylsulfate	3.21	3.27	3.27	3.46	3.33
78	1-butyl-3-methylimidazolium	1-octylsulfate	3.23	2.92	3.36	2.85	3.21
79	1-butyl-3-methylimidazolium	hexafluorophosphate	3.10	3.26	3.45	3.39	3.33
80	1-butyl-3-methylimidazolium	thiocyanate	3.42	3.33	3.24	3.66	3.30
81	1-butyl-3-methylimidazolium	trifluorotris(pentafluoroethyl)phosphate	1.81	2.49	2.05	2.22	1.97
82	1-decyl-3-methylimidazolium	tetrafluoroborate	0.77	1.36	0.93	1.20	0.86
83	1-ethyl-3-methylimidazolium	bis(pentafluoroethyl)phosphinate	2.83	3.76	3.80	3.40	2.84
84	1-ethyl-3-methylimidazolium	1-ethylsulfate	3.93	3.79	3.81	3.94	3.93
85	1-ethyl-3-methylimidazolium	hydrogensulfate	3.99	3.86	3.79	4.21	3.93
86	1-ethyl-3-methylimidazolium	hexafluorophosphate	3.92	3.84	3.90	4.06	3.96
87	1-ethyl-3-methylimidazolium	thiocyanate	4.23	3.92	3.94	4.32	4.09
88	1-ethyl-3-methylimidazolium	toluene-4-sulfonate	3.81	3.67	3.81	3.74	3.87
89	1-ethyl-3-methylimidazolium	trifluorotris(pentafluoroethyl)phosphate	3.23	3.07	2.85	2.88	3.17
90	1-heptyl-3-methylimidazolium	tetrafluoroborate	2.58	2.39	2.20	2.09	2.40
91	1-hexyl-3-methylimidazolium	tetrafluoroborate	2.98	2.69	2.68	2.39	2.88
92	1-hexyl-3-methylimidazolium	hexafluorophosphate	2.91	2.68	2.69	2.80	2.85
93	1-hexyl-3-methylimidazolium	trifluorotris(pentafluoroethyl)phosphate	1.53	1.91	1.33	1.63	1.46
94	ethyl(2-ethoxyethyl)dimethylammonium	bis(trifluoromethylsulfonyl)amide	3.28	3.27	3.12	3.29	3.35
95	1-heptyl-3-methylimidazolium	hexafluorophosphate	2.30	2.38	2.26	2.50	2.51
96	1-butyl-4-methylpyridinium	tetrafluoroborate	2.98	3.16	3.16	2.81	3.04
97	3-methyl-1-octylimidazolium	hexafluorophosphate	1.96	2.08	1.76	2.21	2.08
98	ethyl(3-methoxypropyl)dimethylammonium	bis(trifluoromethylsulfonyl)amide	3.54	3.49	3.27	3.42	3.40
99	1-ethyl-3-methylimidazolium	chloride	3.86	4.13	4.26	4.06	3.96
100	(cyanomethyl)ethyldimethylammonium	bis(trifluoromethylsulfonyl)amide	3.87	3.93	3.70	4.11	3.99
101	1-(2-ethoxyethyl)pyridinium	bromide	4.24	3.80	4.03	3.79	3.71
102	1-butyl-3-methylimidazolium	methanesulfonate	3.51	3.33	3.19	3.44	3.37
103	1-hexyl-3-methylimidazolium	chloride	2.82	2.97	2.84	2.80	2.85
104	1-(ethoxymethyl)pyridinium	bis(trifluoromethylsulfonyl)amide	3.12	3.34	3.20	3.23	3.31
105	1-butyl-3-methylimidazolium	tetrafluoroborate	3.11	3.28	3.26	2.98	3.37
106	1-ethyl-3-methylimidazolium	methylsulfate	4.20	3.85	3.78	4.12	4.15
107	1-(2-hydroxyethyl)pyridinium	bis(trifluoromethylsulfonyl)amide	3.79	3.65	3.55	3.65	3.61
108	1-(3-hydroxypropyl)pyridinium	bis(trifluoromethylsulfonyl)amide	3.55	3.49	3.63	3.62	3.61
109	1-butyl-3-methylimidazolium	toluene-4-sulfonate	3.29	3.09	3.35	3.08	3.25
110	1-(2-ethoxyethyl)-1-methylpiperidinium	bromide	4.31	3.91	4.17	3.96	4.17
111	1-hexyl-3-methylimidazolium	bis(trifluoromethylsulfonyl)amide	2.24	2.22	1.96	2.20	2.23
112	1-decyl-3-methylimidazolium	hexafluorophosphate	1.50	1.35	1.22	1.61	1.15
113	benzyldodecyldimethylammonium	chloride	0.28	0.35	0.71	0.39	0.45
114	ethyl(3-hydroxypropyl)dimethylammonium	bis(trifluoromethylsulfonyl)imide	3.83	3.67	3.66	3.52	3.50
115	Ethyl(2-hydroxyethyl)dimethylammonium	bis(trifluoromethylsulfonyl)imide	3.70	3.93	3.65	3.86	3.85
116	1-(ethoxymethyl)-1-methylpyrrolidinium	bis(trifluoromethylsulfonyl)imide	3.26	3.68	3.41	3.28	3.33
117	(ethoxycarbonylmethyl)ethyldimethylammonium	bis(trifluoromethylsulfonyl)imide	3.53	3.45	3.66	3.65	3.70
118	1-methyl-3-pentylimidazolium	hexafluorophosphate	3.07	2.97	2.68	3.10	3.11
119	1-ethylpyridinium	chloride	4.22	4.14	3.93	4.35	4.29
120	1-(3-hydroxypropyl)-3-methylimidazolium	bis(trifluoromethylsulfonyl)imide	3.66	3.48	3.69	3.47	3.53
121	1-(3-methoxypropyl)-1-methylpyrrolidinium	bis(trifluoromethylsulfonyl)imide	3.40	3.42	3.33	3.31	3.43
122	1-(3-methoxypropyl)-3-methylimidazolium	bis(trifluoromethylsulfonyl)imide	3.34	3.04	3.33	3.03	3.29
123	1-Hexyl-3-ethylimidazolium	bromide	2.01	2.52	2.32	2.55	2.44
124	1-(ethoxymethyl)-3-methylimidazolium	chloride	3.60	4.10	3.64	3.71	3.68
125	Ethyl(3-methoxypropyl)dimethylammonium	bis(trifluoromethylsulfonyl)imide	3.54	3.23	3.08	3.08	3.24
126	1-butyl-3,5-dimethylpyridinium	chloride	3.42	3.23	3.23	2.95	2.99
127	1-(ethoxymethyl)-3-methylimidazolium	bis(trifluoromethylsulfonyl)imide	3.20	3.35	3.26	3.11	3.22

No. 1-93: training set.; No. 94-112: test set. ; No.113-127: validation set.

$$p_C(\sigma_m) = \sum_{i=1}^{k} v_i p_i^G(\sigma_m) \quad (1)$$

Where $p_C(\sigma_m)$ is the surface area with a charge density of σ_m in cation; k is the number of group types; v_i is the frequency of group i ; $p_i^G(\sigma_m)$ is the contribution of group i on the σ-profile of cation at screening charge density of σ_m.

In order to derive $p_i^G(\sigma_m)$, a linear regression is performed for each of the 51 elements using the σ-profile of 828 cations which are calculated by DMol[3]:

$$OF = \min \sum_{i=1}^{828} \left(p_i(\sigma_m) - \sum_{j=1}^{k} Q(i,j) p_i^G(\sigma_m) \right)^2 \quad (2)$$

Where $p_i(\sigma_m)$ is the surface area of cation i with a charge density of σ_m

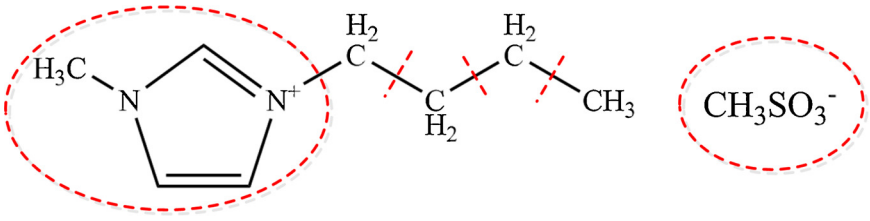


Fig. 2. Group segmentation exemplified for [BMIM][CH_3SO_3].

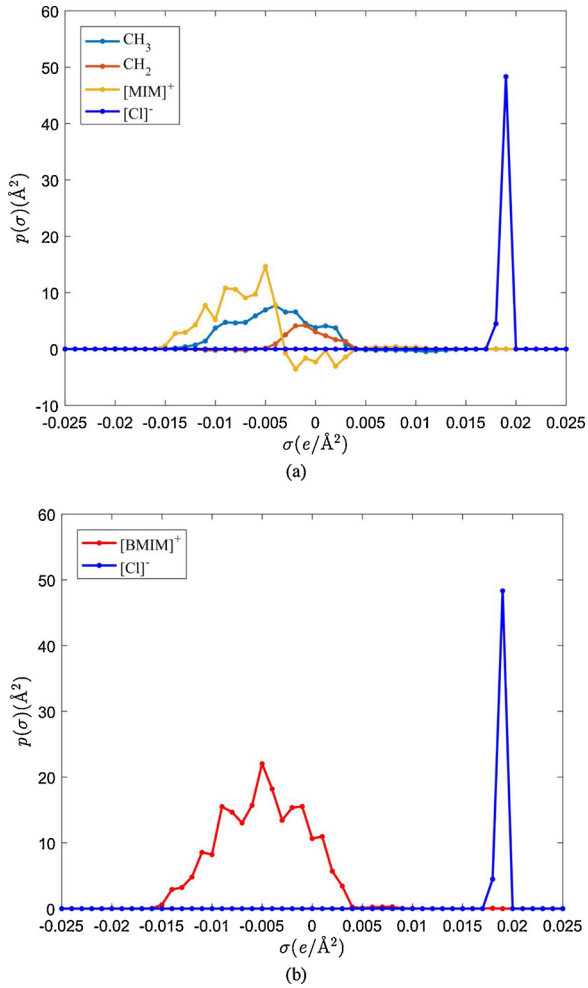


Fig. 3. The σ-profile of (a) groups and (b) total σ-profile of [BMIM][Cl] calculated by GC-COSMO.

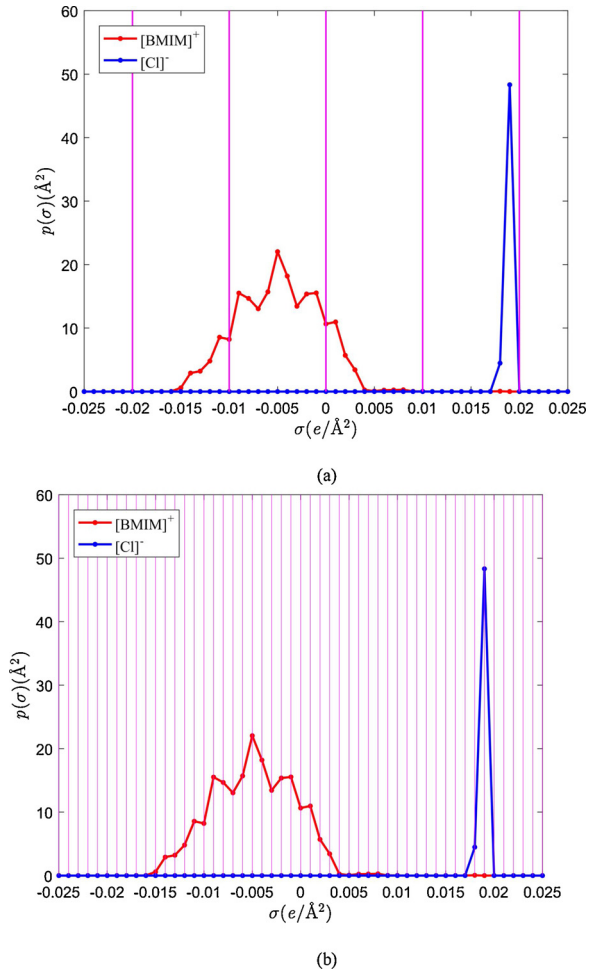


Fig. 4. The segmentation (a) method-1 and (b) method-2 for σ-profile of ILs exemplified for [BMIM][Cl].

calculated by DMol^3; $Q(i, j)$ is the frequency of group j in cation i.

To acquire the σ-profile by DMol^3, the molecular structures of cations and anions are firstly optimized to the lowest energy in the ideal gas phase using the density functional theory (DFT) with VWN-BP functional at the DNP v4.0 basis set (Delley, 2000). After the structural optimization, COSMO files of cations and anions can be acquired by single-point quantum COSMO calculation with the dielectric constant set to infinity. Based on the information in COSMO file, the σ-profile of cations and anions can be obtained.

It is worth mentioning that, the σ-profile of groups is already acquired in our previous work, so in this work, the regression mentioned above is no need to be repeated.

2.3. Descriptor

After generating the σ-profile for all the cations and anions in the dataset, two segmentation methods are used to calculate the descriptors for developing the QSAR model. For the first method (method-1), as seen from Fig. 4a, the σ-profile of both cation and anion are dived into 6 parts: $s_{\sigma(-0.025\sim-0.02)}$, $s_{\sigma(-0.02\sim-0.01)}$, $s_{\sigma(-0.01\sim0)}$, $s_{\sigma(0\sim0.01)}$, $s_{\sigma(0.01\sim0.02)}$, $s_{\sigma(0.02\sim0.025)}$. The area under each region is calculated and their numerical value is treated as the descriptor, and thus there are 12 descriptors altogether.

For the second method (method-2), the descriptor is the surface area with a charge density of σ_m ($p(\sigma_m)$). Because the σ-profile for both cation and anion are defined as a vector of 51 elements (Fig. 2), there are 102 descriptors in total. For the cation, they are denoted as $s_{\sigma-0.025C}$, $s_{\sigma-0.024C}$, …, $s_{\sigma0.024C}$, $s_{\sigma0.025C}$, and for the anion as $s_{\sigma-0.025A}$, $s_{\sigma-0.024A}$, …, $s_{\sigma0.024A}$, $s_{\sigma0.025A}$.

2.4. ERM

It is already been proved that molecular descriptors play vital roles in building models (Zhao et al., 2015). In this work, ERM (Mercader et al., 2008) is used to find out the best subset **d** from the pool of descriptors **D** with **d** $\ll$ **D** which reaches the minimal standard deviation S of MLR model.

$$S = \frac{1}{(N - d - 1)} \sum_{i=1}^{N} res_i^2 \tag{3}$$

Where N is the number of IL in the training set; res_i is the residual for IL i.

ERM is a modified version of the replacement method (Duchowicz et al., 2006), it exhibits less propensity for remaining in local minima and at the same time is less dependent on the initial solution. Moreover, it requires a smaller number of linear regressions than a time-consuming Full Search (FS) method while obtaining identical results. This technique approaches the minimum of S by judiciously taking into account the relative errors of the coefficients of the least-squares model given by a set of d descriptors. The ERM gives models with better statistical parameters than the Forward Stepwise Regression procedure (Kumar et al., 2011; Simon and Abdelmalek, 2012) and the more elaborated Genetic Algorithms (Jouyban et al., 2014; Mercader et al., 2010). It has been utilized with satisfactory results in many QSAR/

QSPR reports (Abooali and Sobati, 2014; Rybka et al., 2014; Sun et al., 2009).

2.5. MLR

After acquiring the optimal subset of descriptors, MLR is applied to establish the linear relationship of the chosen descriptors and the toxicity of ILs, the generalized expression for the MLR can be written as follows,

$$\log_{10}(EC_{50}) = c_0 + \sum_{i=1}^{p} x_i * c_i \quad (4)$$

Where c_0 is the constant term, and c_i is the estimated coefficient of the corresponding descriptor x_i; p denotes the number of descriptors.

The sign of the coefficient of Eq. 4 can help us to understand the influence of each descriptor on the toxicity. The positive value means parameters are positively-related to the toxicity while the negative values mean parameters are negatively-related to the toxicity. It should be noted that a lower logarithmic value corresponds to the higher toxicity of ILs. Moreover, the importance of every descriptor can be illustrated from the t and p value of the MLR model.

2.6. MLP

MLP method was used to build the non-linear QSAR model by the Neural Network Toolbox in Matlab (R2016b version). A MLP is a class of feedforward artificial neural network (*NN*), it consists of three layers of nodes: an input layer, a hidden layer and an output layer. The input neuron number equal to the number of descriptors while the output neuron number is one in this work. The hidden neuron number (*HNN*) is related to the converging performance of the output error function during the training process. Too few *HNN* values would hamper the learning capability of the *NN*, while too many can cause over-fitting or memorization of the learning sample. Each neuron receives information of all the neurons from the previous layer, and every connection is controlled by parameters called weights, which are optimized by Back-propagation (BP) training function. During the training procedure, learning coefficient (*LC*), which define the learning capability of a neural network, is used to control the degree at which connection weights are modified in the learning phase. In order to design the best MLP model with the minimum *MSE* for the training set, the parameters *HNN* and *LC* are optimized. The robustness of the final model is tested by Leave-One-Out (LOO) cross validation and the external validation.

3. Results and discussion

ERM is used to search the best subset of descriptors for developing the QSAR models. The contribution of different groups to the toxicity of ILs are calculated and discussed to validate the reliability of the chosen descriptors. In order to design the best MLP model, the hidden neuron number and the learning coefficient are optimized to minimize the *MSE* for the training set. To assess the robustness of both linear and non-linear models and avoid overfitting, internal and external validation are performed. The internal validation using the LOO cross-validation technique, while the external validation predicting the logEC50 value of 15 new ILs which are excluded from the training and test set.

3.1. Descriptor selection and validation

To determine the optimum number of descriptors for the two segmentation methods mentioned above, a variety of subset sizes are investigated. The best-correlations with experimental toxicity (logEC50) are selected on the basis of the *MSE* and R^2 of train and test set using MLR model (Appendix A).

As shown in Fig. 5a, for the training set of method-1, R^2 increases while *MSE* decreases with the increasing number of the descriptors.

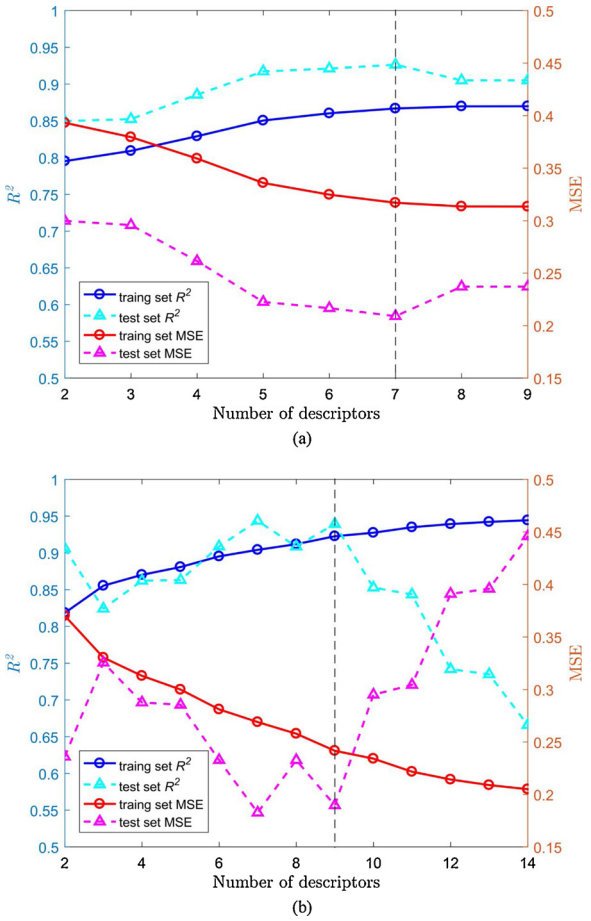


Fig. 5. MSE and R^2 of the (a) MLR-1 model and (b) MLR-2 model versus the number of descriptors for the training and test sets.

When the number of the descriptor increased to 7, the change for both R^2 and *MSE* can be neglected. In the case of the test set, R^2 begins to decrease while *MSE* begins to increase when the number of descriptors over 7. Thus, the optimal subset is obtained when the number of descriptors is 7, and the coefficient of the final MLR model for method-1 (MLR-1) are listed in Table 2 and ranked by p value in ascending order. The lower the value of p value means the more important of the descriptor. It can be seen from Table 2 that cations have the major effect on the toxicity of ILs since x_2 to x_4 are all cation-related items and their p value are close to that of the anion-related item x_1.

For method-2, as shown in Fig. 5b, the same variation tendency of R^2 and *MSE* of training and test set can be found when the number of descriptors is over 9. The coefficient of the final MLR model for model-2 (MLR-2) is listed in Table 3. It is found that p value of the descriptor of $S_{\sigma 0.003C}$ is much lower than other descriptors, which means it has a dominant effect on the toxicity of ILs in method-2. Moreover, x_1 to x_3

Table 2
The results of MLR-1 model

	Description*	Coefficient	t value	p value
$x0$	Constant	7.6896	11.7900	1.42E-19
$x1$	$S_{\sigma(-0.01A-0A)}$	−0.0057	−5.8393	9.34E-08
$x2$	$S_{\sigma(-0.01C-0C)}$	−0.0163	−5.2013	1.35E-06
$x3$	$S_{\sigma(0.01C-0.02C)}$	0.0674	4.6482	1.21E-05
$x4$	$S_{\sigma(-0.02C--0.01C)}$	−0.0362	−4.3050	4.45E-05
$x5$	$S_{\sigma(0A-0.01A)}$	−0.0083	−2.4006	1.86E-02
$x6$	$S_{\sigma(0C-0.01C)}$	−0.0109	−2.0967	3.90E-02
$x7$	$S_{\sigma(0.02C-0.025C)}$	−12.2490	−2.0463	4.38E-02

* subscripts A and C mean anions and cations, respectively.

Table 3
The results of MLR-2 model

	Description*	Coefficient	*t* value	*p* value
*x*0	Constant	5.4003	31.3080	9.90E-48
*x*1	$S_{\sigma 0.003C}$	−0.2353	−24.6110	6.80E-40
*x*2	$S_{\sigma\text{-}0.004C}$	−0.0710	−8.2541	2.03E-12
*x*3	$S_{\sigma 0.004C}$	0.2324	6.3200	1.24E-08
*x*4	$S_{\sigma 0.012A}$	−0.0202	−5.3679	7.09E-07
*x*5	$S_{\sigma 0.019C}$	20.7320	5.1204	1.94E-06
*x*6	$S_{\sigma 0.003A}$	−0.0099	−4.4959	2.23E-05
*x*7	$S_{\sigma 0.013A}$	0.0289	4.3570	3.75E-05
*x*8	$S_{\sigma 0.002A}$	−0.0082	−4.2382	5.81E-05
*x*9	$S_{\sigma\text{-}0.003A}$	−0.0307	−3.7222	3.59E-04

* subscripts A and C mean anions and cations, respectively.

are all cation-related descriptors, it again proves that cations have a remarkable effect on the toxicity of ILs. Additionally, the *p* values in Table 2 and 3 are all lower than 0.05, which means all the selected descriptors have significant contributions to the toxicity of ILs.

In order to validate the reliability of the selected descriptors, the contribution of diverse groups to the toxicity of ILs are systematically analysed by GC-COSMO method and MLR model. The contribution of each group to the toxicity is calculated using the selected descriptors and the corresponding parameters listed in Table 2 and 3. For example, the cation-related descriptors x_2, x_3, x_4, x_6 and x_7 in method-1 (Table 2) for CH_3 are 56.12, -1.40, 4.72, 9.37 and -0.03, respectively. Therefore, the contribution of CH_3 can be calculated as $(56.12 \times -0.0162) + (-1.4 \times 0.0674) + (4.72 \times -0.0362)$. By this method, the contribution of different groups can be calculated, the results being listed in Table 4. As it can be seen, the contributions of CH_2 are -0.37 and -0.30 calculated by method-1 and method-2, respectively. This indicates that increasing the number of CH_2 will lower the log*EC*50 value and make IL more toxic towards the IPC-81. This can be explained by the fact that longer alkyl chains may be incorporated into the polar head groups of the phospholipid bilayer, which is the major structure of membranes, thus the cell membrane can be easily damaged (Singh et al., 2014). To investigate the influence of the presence of oxygen on the toxicity, the contribution of OH and OCH_2 are calculated and compared with CH_3 and $(CH_2)_2$, respectively. As it can be seen from Table 4, the contribution value of all oxygenated groups are higher than the alkyl groups and consequently result in higher log*EC*50 value, which indicates that introduction of oxygen groups into the alkyl side chain significantly reduced the toxicity of ILs (Tot et al., 2018).

$$+ (9.37 \times -0.0109) + (-0.03 \times -12.2486) = -0.95$$

The influence of the cation skeleton MIM, MPYO and MPI on the toxicity of ILs is also investigated, the contribution value calculated by method-2 is presented in the following order MIM < MPYO < MPI. The imidazolium ILs are the most toxic may be due to the specific character of the imidazolim head group including the hydrogen bonding (Smirnova and Safonova, 2010). The result is consistent with experimental data where the log*EC*50 value for $[C_4MIM][Br]$, $[C_4MPYO][Br]$ and $[C_4MPI][Br]$ are 3.43, 3.77 and 4.03, respectively. By contrast, the results from method-1 is inconsistent with the experimental data. Considering the performance of MLR-2 is significantly better than MLR-1, method-2 is more suitable for building the QSAR model for the prediction of toxicity of ILs.

Table 4
The contribution of different groups to the toxicity of ILs.

Categories	Groups	method-1	method-2
Substituent	CH_3	−0.95	−0.94
	CH_2	−0.37	−0.30
	OH	−0.18	−0.32
	OCH_2	0.14	0.04
Cation skeleton	MPI	−1.34	0.34
	MPYO	−1.12	−0.01
	MIM	−1.30	−0.17
Anion	Cl	0.00	0.00
	TOS	−0.87	−0.32
	$MDEGSO_4$	−0.88	−0.47
	Tf_2N	−1.52	−0.60
	eFAP	−2.15	−1.17

Concerning to the anion effect in ILs toxicity, five anions with the same cation $[C_4MIM]^+$ are compared ([Cl]- (3.55), [TOS]- (3.29), $[MDEGSO_4]$- (3.15), $[Tf_2N]$- (2.68), [eFAP]- (1.81)). As seen from Table 4, for both methods, the contribution value of anions to the toxicity of ILs follows the order: [Cl]- > [TOS]- > [MDEGSO4]- > $[Tf_2N]$- > [eFAP]-. The toxicity of the fluorine-containing anions is obviously higher than other kinds of anion (Costa et al., 2015), which is consistent with the experimental data.

These findings validate that the selected descriptors are highly correlated with the toxicity of ILs, and it is reasonable to using them to develop the QSAR models.

3.2. The QSAR models based on method-1

The corresponding plots of experimental data versus calculated values by MLR-1 and MLP-1 (MLP model for method-1, *HNN* = 6 and *Lc* = 0.0065) are presented in Fig. 6 and the statistical parameters are listed in Table 6. It can be seen that good correlation relationship (R^2 = 0.867 and 0.959 for the training set, respectively) can be found

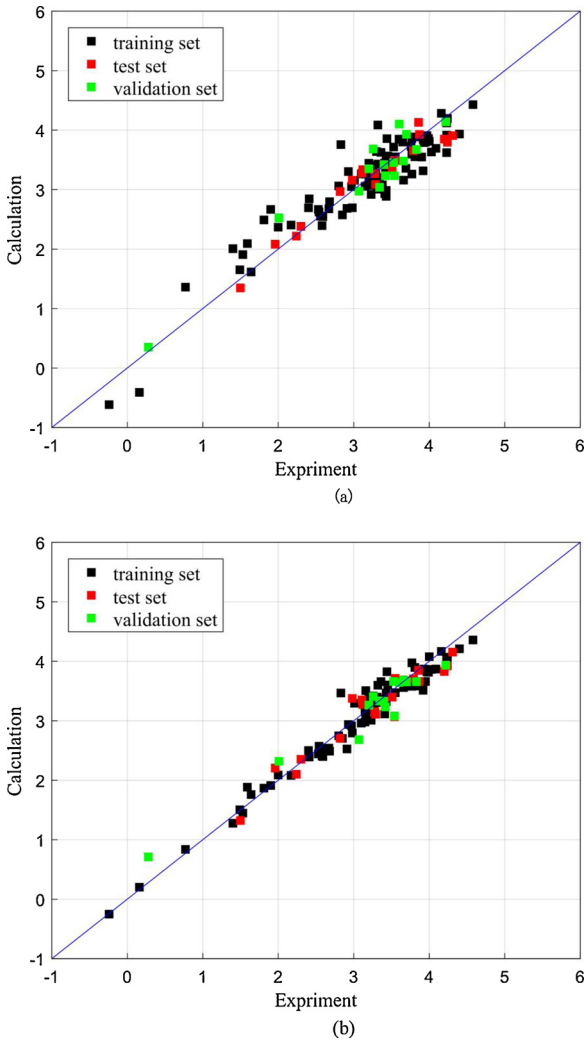


Fig. 6. Calculated versus experimental toxicity values calculated by: (a) MLR-1 (b) MLP-1.

Table 5
The MSE and R2 of LOO Cross-Validation for training set.

	MLR-1	MLP-1	MLR-2	MLR-2
MSE_{CV}	0.1004	0.0423	0.0583	0.0187
R^2_{CV}	0.8669	0.9439	0.9228	0.9753

Table 6
Comparisons of the statistical parameters by different QSAR models.

Model	Dataset	No.	R^2	R^2_{adjust}	AARD%	MSE	RMSE
MLR-1	Training	93	0.867	0.856	14.357	0.101	0.317
	Test	19	0.926	0.879	5.235	0.044	0.209
	Total	112	0.875	0.867	12.810	0.091	0.302
	Validation	15	0.914	0.827	8.551	0.071	0.267
MLP-1	Training	93	0.959	0.956	4.663	0.031	0.176
	Test	19	0.913	0.857	6.385	0.052	0.228
	Total	112	0.953	0.950	4.955	0.034	0.186
	Validation	15	0.932	0.863	15.157	0.056	0.238
MLR-2	Training	93	0.923	0.914	9.628	0.058	0.242
	Test	19	0.939	0.879	4.949	0.036	0.190
	Total	112	0.925	0.918	8.835	0.055	0.234
	Validation	15	0.915	0.763	9.129	0.070	0.264
MLP-2	Training	93	0.975	0.973	4.493	0.019	0.137
	Test	19	0.938	0.876	5.091	0.037	0.192
	Total	112	0.970	0.968	4.595	0.022	0.147
	Validation	15	0.944	0.844	9.047	0.046	0.214

between the chosen descriptors and the toxicity value of ILs. The satisfactory results of cross-validation (Table 5) indicate the developed model is not over fitted or a result of by-chance. In terms of external validation, the R^2 and *MSE* are all close to the results of the training and the test set, which confirmed the predictive ability of the proposed models.

To define the application domain, Williams plot of the MLR-1 and MLP-1 are presented in Fig. 7. It can be seen that the majority of ILs are located within the application domain (Appendix B) and are predicted accurately, which further confirmed the reliability of the prediction models. The *h* value of 1-hexadecyl-3-methylimidazolium chloride (23, -0.24), benzyltetradecyldimethylammonium chloride (64, 0.16) and benzyldodecyldimethylammonium chloride (113, 0.28) are greater than the threshold leverage value h^* and the standized residuals of these three ILs are also higher than 3. This is because these ILs all have very long alkyl groups which are different from other ILs in the training dataset. Moreover, the prediction error of GC-COSMO will be slightly increased when it comes to ILs have an extremely long alkyl chain.

3.3. The QSAR models based on method-2

The experimental data versus calculated values by MLR-2 and MLP-2 (MLP model for method-2 ($NN = 8$ and $Lc = 0.0014$) are presented in Fig. 8. Compared to the results of method-1, better correlation relationship can be found with R^2 for the training set of 0.923 and 0.975, respectively. The LOO cross-validation ($R_{LOO}{}^2$ and MSE_{LOO} are 0.923 and 0.975, respectively) and external validation (R^2 and *MSE* are 0.915 and 0.944, respectively) confirmed the reliability of the QSAR models based on method-2.

Williams plot of the MLR-2 and MLP-2 are given in Fig. 9. It can be seen that the standized residual of compounds with long alky chain (23, 64 and 113) are still greater than 3. In terms of the leverage value, the *h* of compounds 1-benzyl-3-methylimidazolium tetrafluoroborate (3, 2.97) and 1-butyl-3-methylimidazolium 2-(2-methoxyethoxy)ethylsulfate (6, 3.15) are higher than h^*. This is because the ILs containing the benzyl group or 2-(2-methoxyethoxy)ethylsulfate anion are different from other ILs in the training set. Considering their low standized residual, compounds 3 and 6 can be considered as structurally influential materials in the dataset (Ma et al., 2015).

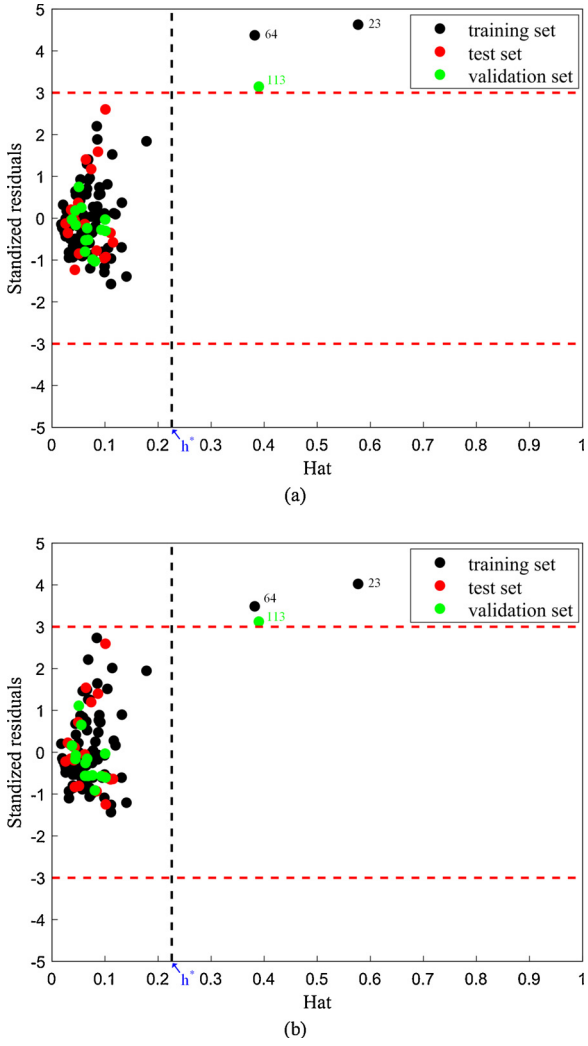


Fig. 7. Williams plot of the training, test and validation sets : (a) MLR-1 and (b) MLP-1.

3.4. Comparisons

Favorable results are obtained for four QSAR models with high R^2 and low *MSE* values (Table 6). The performance of the four QSAR models can be ranked as following order: MLP-2 > MLP-1 > MLR-2 > MLR-1. The performance of the nonlinear model based on the second segmentation method (MLP-2) is better than others, the R^2 values of the training set, test set and validation set are 0.975, 0.938 and 0.970, respectively. It can be seen from Table 6, the second segmentation method is more suitable for generating the descriptors for the prediction of toxicity of ILs. Furthermore, the nonlinear model MLP exhibits better results compared to the linear model MLR.

The comparisons of the QSAR models in the literature developed for prediction of the toxicity of ILs towards IPC-81 are summarized in Table 7. It can be seen that, in general, the best model MLP-2 developed in this work ($R^2 = 0.975$ for training set) is better than the most of models in the literature. However, the results of MLP-2 do not show great improvement compare to the work of Cao et al. (2018) using a similar dataset ($R^2 = 0.974$ for training set), and inferior to the models presented by Torrecilla et al. (2010) and Fatemi and Izadiyan (2011) with $R^2 = 0.996$ and $R^2 = 0.99$, respectively. Compared to the models also using σ-profile as the descriptor (Torrecilla et al., 2010; Cao et al., 2018), the method used in this work is more efficient because the time-consuming quantum mechanical calculations for the σ-profile can be avoided. Moreover, the influence of every group on the toxicity of ILs

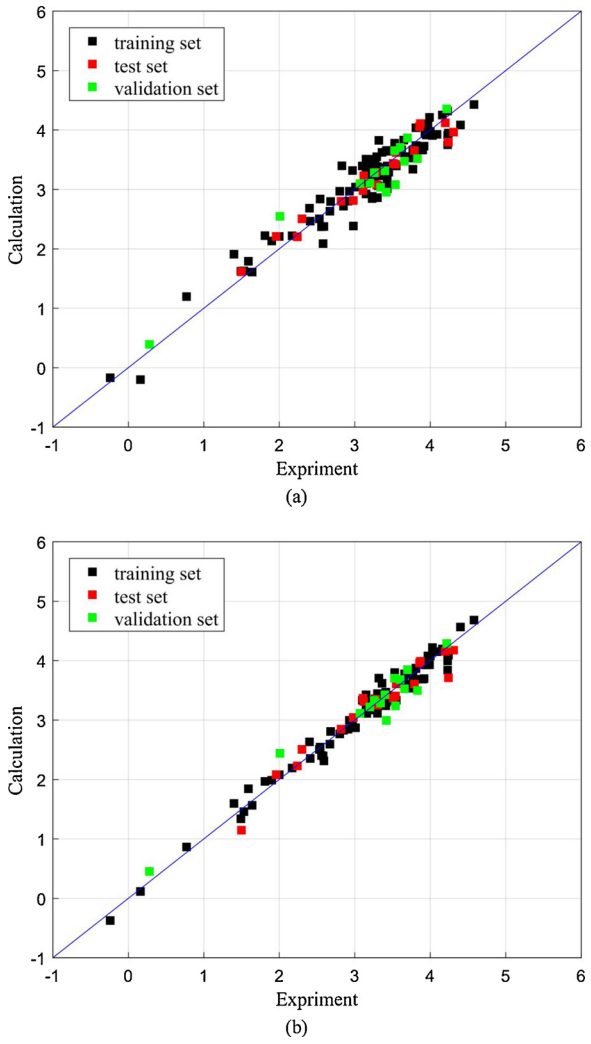


Fig. 8. Calculated versus experimental toxicity values calculated by: (a) MLR-2 (b) MLP-2.

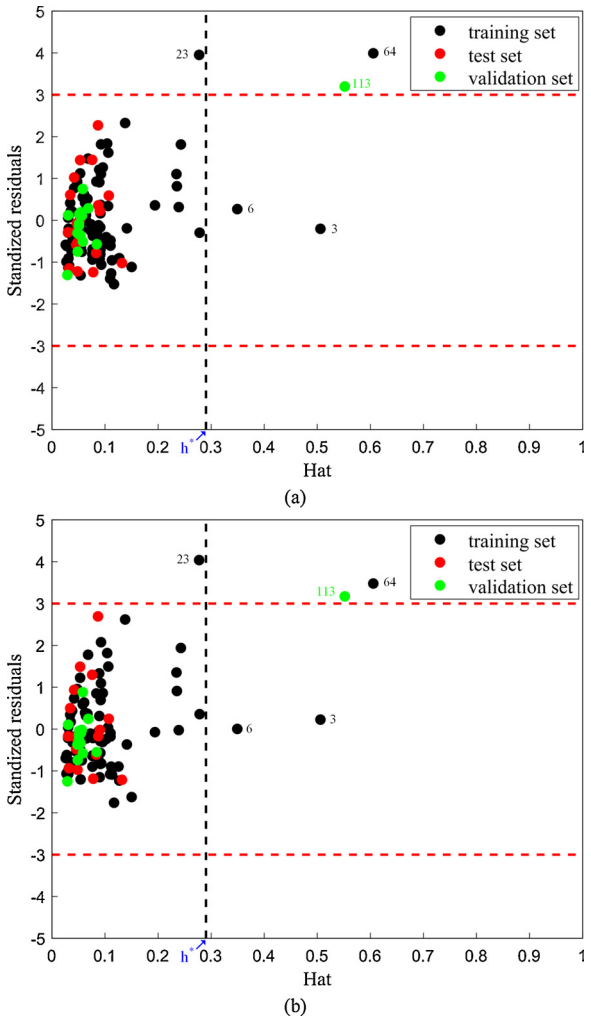


Fig. 9. Williams plot of the training, test and validation sets : (a) MLR-2 and (b) MLP-2.

can be evaluate by the σ-profile of every group. Compared to the model developed by Fatemi and Izadiyan (2011) using GATEWAY (GEometry, Topology and Atom-Weights AssemblY) descriptors, the model presented in this work is more simple, the descriptors can be easily acquired by GC-COSMO method. Another advantage of the presented QSPR model is that since it is GC-based it can be directly used for CAILD.

4. Conclusion

In this work, the σ-profile of 127 ILs are calculated by GC-COSMO method and the corresponding descriptors are acquired by two segmentation methods. In order to acquire the optimal subset of the descriptors, an algorithm called EMR is used, and the best descriptor number for method-1 and method-2 are 7 and 9, respectively. The reliability of the selected descriptors is validated by detailed analysis of the relationship between the structure and the toxicity of ILs and cation is found to have a major effect on the toxicity of ILs. Based on the chosen descriptors, linear and nonlinear QSAR models are established to estimate the toxicity of 127 ILs towards IPC-81. The LOO cross-validation together with the external validation confirmed that all the presented model are reliable and not overt fitter. Among the four proposed QSAR models, the nonlinear model based on the second segmentation method (MLR-2) yielded the best toxicity-structure relationship with $R^2 = 0.975$, $MSE = 0.019$ for the training set and $R^2 = 0.938$, $MSE = 0.037$ for the test set. Compared to other QSAR models in the literature, the QSAR method developed in this work is more efficient, and moreover it can be used to design green ILs with low toxicity by CAILD method.

Declaration of Competing Interest

The authors declare that they have no competing interests.

CRediT authorship contribution statement

Daili Peng: Methodology, Software, Validation, Conceptualization, Formal analysis, Writing - original draft, Visualization, Funding acquisition, Data curation. **Francesco Picchioni:** Conceptualization, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgements

We are grateful for the financial support provided by the China Scholarship Council (No. 201606740070).

Table 7
Summary of published QSAR models for predicting the toxicity of ILs to the IPC-81.

Num. of ILs	Method	Num. of descriptors	Descriptors type	R^2	Software used	Ref.
74	LR (linear regression)	1	Lipophilicity parameters	0.78	R software	Ranke et al. (2007)
153	MLR	12	Elemental composition and molecular weights	0.867*	Matlab and Statgraphics Plus	Torrecilla et al. (2009)
	RB (radial-basis function)			0.861*		
	MLP			0.982*		
105	MLR	10	σ-profile from COSMO-RS	0.9	Matlab, Gaussian and SPSS	Torrecilla et al. (2010)
	MLP			0.996		
227	MLR	5	GATEWAY (GEometry, Topology and Atom-Weights AssemblY) descriptors	0.92	ChemSketch, HyperChem, Dragon and STATISTICA	Fatemi and Izadiyan (2011)
	MLP			0.99		
100	MLR	4	GATEWAY descriptors	0.918	Matlab and CODESSA	Zhao et al. (2014)
	SVM (support vector machine)			0.959		
55	MLR	10	Group contribution descriptors	0.9184	SPSS	Peric et al. (2015)
304	MLR	5	GATEWAY descriptors	0.772	ChemSketch, MOPAC, Dragon and QSARINs	Sosnowska et al. (2017)
269	PLS (least squares)	140	GRINDs (Alignment free GRid-INdependent Descriptors)	0.86	MATLAB, MarvinSketch, HyperChem, Dragon, and Pentacle	Farahani et al. (2018)
	SVR (support vector regression)			0.89		
119	MLR	8	The SEP (electrostatic potential surface area) and σ-profile from COSMO-RS	0.93	Matlab, Gaussian and Mutiwfn	Cao et al. (2018)
	SVM			0.951		
	ELM (extreme learning machine)			0.974		
127	MLR-1	7	σ-profile from GC-COSMO	0.867	Matlab	This work
	MLP-1			0.959		
	MLR-2	9		0.923		
	MLP-2			0.975		

* Means R^2 for the test set.

Appendix A. Evaluation

The performance of the QSAR model is measured by different metrics, i.e. squared correlation coefficient (R^2), adjusted squared correlation coefficient ($R_{adj}{}^2$), average absolute relative deviation ($AARD\%$), mean square error (MSE), root mean square error ($RMSE$), the squared correlation coefficient ($R_{LOO}{}^2$) and mean square error (MSE_{LOO}) of Leave-One-Out cross validation for training set, the corresponding equations are listed below,

$$R^2 = \frac{\sum_{i=1}^{N}(y_i^{cal} - \bar{y}_{train})^2 - \sum_{i=1}^{N}(y_i^{cal} - y_i^{exp})^2}{\sum_{i=1}^{N}(y_i^{cal} - \bar{y}_{train})^2} \tag{A.1}$$

$$R_{adj}{}^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1} \tag{A.2}$$

$$AARD(\%) = 100 \times \sum_{i=1}^{N} \left| \frac{y_i^{cal} - y_i^{exp}}{y_i^{exp}} \right| / N \tag{A.3}$$

$$MSE = \sum_{i=1}^{N}(y_i^{cal} - y_i^{exp})^2 / N \tag{A.4}$$

$$RMSE = \sqrt{\sum_{i=1}^{N}(y_i^{cal} - y_i^{exp})^2 / N} \tag{A.5}$$

$$R_{LOO}{}^2 = \frac{\sum_{i=1}^{Nt} Rout_i{}^2}{Nt} \tag{A.6}$$

$$MSE_{LOO} = \frac{\sum_{i=1}^{Nt} MSEout_i}{Nt} \tag{A.7}$$

where y_i^{cal} is the calculation value of IL i, while y_i^{exp} is the experimental value. N is the number of IL in the data set. $\bar{y}_{train}$ and N_t is the mean value of experimental log$EC50$ and the number of ILs in training set, respectively. $Rout_i{}^2$ and $MSEout_i$ denote the R^2 and MSE after leaving the i th IL out of the training set, respectively.

Appendix B. Application domain

The application domain is a theoretical spatial region defined by the values of molecular descriptors and the modelled response. In the presented study, the application domain was verified by using the leverages (Williams plot) (Gramatica, 2007) and standardization approach (Roy et al., 2015). The leverage value of compound i (h_i) is calculated from the descriptors matrix (X),

$$h_i = X_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_i \tag{B.1}$$

where x_i is a row vector of descriptors for compound i and X is the matrix of descriptors for the training set.

The boundary of the application domain is defined by the critical value of leverage, h^* and the values of the standardized residuals differing by more than ± 3 standard deviation units. The critical value of leverage can be calculated as

$$h^* = 3p/n \tag{B.2}$$

where p is the number of variables used in the model and n is the number of the training data.

Those values of h_i higher than threshold value h^* mean that the structure of a compound significantly differs from other compounds in the training data.

References

Abooali, D., Sobati, M.A., 2014. Novel method for prediction of normal boiling point and enthalpy of vaporization at normal boiling point of pure refrigerants: A QSPR approach. Int. J. Refrig. 40, 282–293. https://doi.org/10.1016/j.ijrefrig.2013.12.007.

Bates, E.D., Mayton, R.D., Ntai, I., Davis, J.H., 2002. CO2 capture by a task-specific ionic liquid. J. Am. Chem. Soc. 124, 926–927. https://doi.org/10.1021/ja017593d.

Cao, L., Zhu, P., Zhao, Y., Zhao, J., 2018. Using machine learning and quantum chemistry descriptors to predict the toxicity of ionic liquids. J. Hazard. Mater. 352, 17–26. https://doi.org/10.1016/j.jhazmat.2018.03.025.

Chen, B.-K., Liang, M.-J., Wu, T.-Y., Wang, H.P., 2013. A high correlate and simplified QSPR for viscosity of imidazolium-based ionic liquids. Fluid Phase Equilib. 350, 37–42. https://doi.org/10.1016/j.fluid.2013.04.009.

Chen, K., Lin, W., Yu, X., Luo, X., Ding, F., He, X., Li, H., Wang, C., 2015. Designing of anion-functionalized ionic liquids for efficient capture of SO $_2$ from flue gas. AIChE J. 61, 2028–2034. https://doi.org/10.1002/aic.14793.

Cho, C.-W., Ranke, J., Arning, J., Thöming, J., Preiss, U., Jungnickel, C., Diedenhofen, M., Krossing, I., Stolte, S., 2013. *In silico* modelling for predicting the cationic hydrophobicity and cytotoxicity of ionic liquids towards the *Leukemia* rat cell line, *Vibrio fischeri* and *Scenedesmus vacuolatus* based on molecular interaction potentials of ions. SAR QSAR Environ. Res. 24, 863–882. https://doi.org/10.1080/1062936X.2013.821092.

Costa, S.P.F., Pinto, P.C.A.G., Lapa, R.A.S., Saraiva, M.L.M.F.S., 2015. Toxicity assessment of ionic liquids with Vibrio fischeri: An alternative fully automated methodology. J. Hazard. Mater. 284, 136–142. https://doi.org/10.1016/j.jhazmat.2014.10.049.

Delley, B., 2000. From molecules to solids with the DMol3 approach. J. Chem. Phys. 113, 7756–7764. https://doi.org/10.1063/1.1316015.

Duchowicz, P.R., Castro, E.A., Fernández, F.M., 2006. Alternative Algorithm for the Search of an Optimal Set of Descriptors in QSAR-QSPR Studies. MATCH Commun. Math. Comput. Chem 55, 179–192.

Eshetu, G.G., Armand, M., Ohno, H., Scrosati, B., Passerini, S., 2016. Ionic liquids as tailored media for the synthesis and processing of energy conversion materials. Energy Environ. Sci. 9, 49–61. https://doi.org/10.1039/C5EE02284C.

Farahani, S.R., Sohrabi, M.R., Ghasemi, J.B., 2018. A detailed structural study of cytotoxicity effect of ionic liquids on the leukemia rat cell line IPC-81 by three dimensional quantitative structure toxicity relationship. Ecotoxicol. Environ. Saf. 158, 256–265. https://doi.org/10.1016/j.ecoenv.2018.04.040.

Fatemi, M.H., Izadiyan, P., 2011. Cytotoxicity estimation of ionic liquids based on their effective structural features. Chemosphere 84, 553–563. https://doi.org/10.1016/j.chemosphere.2011.04.021.

García-Lorenzo, A., Tojo, E., Tojo, J., Teijeira, M., Rodríguez-Berrocal, F.J., González, M.P., Martínez-Zorzano, V.S., 2008. Cytotoxicity of selected imidazolium-derived ionic liquids in the human Caco-2 cell line. Sub-structural toxicological interpretation through a QSAR study. Green Chem. 10, 508. https://doi.org/10.1039/b718860a.

Ghanem, O. Ben, Mutalib, M.I.A., Lévêque, J.-M., El-Harbawi, M., 2017. Development of QSAR model to predict the ecotoxicity of Vibrio fischeri using COSMO-RS descriptors. Chemosphere 170, 242–250. https://doi.org/10.1016/j.chemosphere.2016.12.003.

Gharagheizi, F., Ilani-Kashkouli, P., Mohammadi, A.H., 2012. Group contribution model for estimation of surface tension of ionic liquids. Chem. Eng. Sci. 78, 204–208.

https://doi.org/10.1016/J.CES.2012.05.008.
Gramatica, P., 2007. Principles of QSAR models validation: internal and external. QSAR Comb. Sci. 26, 694–701. https://doi.org/10.1002/qsar.200610151.
Hossain, M.I., Samir, B.B., El-Harbawi, M., Masri, A.N., Mutalib, M.I.A., Hefter, G., Yin, C.-Y., 2011. Development of a novel mathematical model using a group contribution method for prediction of ionic liquid toxicities. Chemosphere 85, 990–994. https://doi.org/10.1016/j.chemosphere.2011.06.088.
Huang, Y., Dong, H., Zhang, X., Li, C., Zhang, S., 2013. A new fragment contribution-corresponding states method for physicochemical properties prediction of ionic liquids. AIChE J. 59, 1348–1359. https://doi.org/10.1002/aic.13910.
Jouyban, A., Shayanfar, A., Ghafourian, T., Acree, W.E., 2014. Solubility prediction of pharmaceuticals in dioxane + water mixtures at various temperatures: Effects of different descriptors and feature selection methods. J. Mol. Liq. 195, 125–131. https://doi.org/10.1016/j.molliq.2014.02.012.
Kumar, S., Singh, V., Tiwari, M., 2011. QSAR modeling of the inhibition of reverse transcriptase enzyme with benzimidazolone analogs. Med Chem Res 20, 1530–1541. https://doi.org/10.1007/s00044-010-9406-2.
Lazzús, J.A., 2012. A group contribution method to predict the melting point of ionic liquids. Fluid Phase Equilib. 313, 1–6. https://doi.org/10.1016/J.FLUID.2011.09.018.
Lazzús, J.A., Pulgar-Villarroel, G., 2015. A group contribution method to estimate the viscosity of ionic liquids at different temperatures. J. Mol. Liq. 209, 161–168. https://doi.org/10.1016/J.MOLLIQ.2015.05.030.
Luis, P., Garea, A., Irabien, A., 2010. Quantitative structure–activity relationships (QSARs) to estimate ionic liquids ecotoxicity EC50 (Vibrio fischeri). J. Mol. Liq. 152, 28–33. https://doi.org/10.1016/J.MOLLIQ.2009.12.008.
Luis, P., Ortiz, I., Aldaco, R., Irabien, A., 2007. A novel group contribution method in the development of a QSAR for predicting the toxicity (Vibrio fischeri EC 50) of ionic liquids. Ecotoxicol. Environ. Saf. 67, 423–429. https://doi.org/10.1016/j.ecoenv.2006.06.010.
Lyu, Z., Zhou, T., Chen, L., Ye, Y., Sundmacher, K., Qi, Z., 2014. Reprint of: Simulation based ionic liquid screening for benzene-cyclohexane extractive separation. Chem. Eng. Sci. 115, 186–194. https://doi.org/10.1016/j.ces.2014.05.032.
Ma, S., Lv, M., Deng, F., Zhang, X., Zhai, H., Lv, W., 2015. Predicting the ecotoxicity of ionic liquids towards Vibrio fischeri using genetic function approximation and least squares support vector machine. J. Hazard. Mater. 283, 591–598. https://doi.org/10.1016/J.JHAZMAT.2014.10.011.
Mercader, A.G., Duchowicz, P.R., Fernández, F.M., Castro, E.A., 2010. Replacement Method and Enhanced Replacement Method Versus the Genetic Algorithm Approach for the Selection of Molecular Descriptors in QSPR/QSAR Theories. J. Chem. Inf. Model. 50, 1542–1548. https://doi.org/10.1021/ci100103r.
Mercader, A.G., Duchowicz, P.R., Fernández, F.M., Castro, E.A., 2008. Modified and enhanced replacement method for the selection of molecular descriptors in QSAR and QSPR theories. Chemom. Intell. Lab. Syst. 92, 138–144. https://doi.org/10.1016/J.CHEMOLAB.2008.02.005.
Mirkhani, S.A., Gharagheizi, F., Ilani-Kashkouli, P., Farahani, N., 2012. An accurate model for the prediction of the glass transition temperature of ammonium based ionic liquids: A QSPR approach. Fluid Phase Equilib. 324, 50–63. https://doi.org/10.1016/J.FLUID.2012.03.024.
Mullins, E., Oldland, R., Liu, Y.A., Wang, S., Sandler, S.I., Chen, C.-C., Zwolak, M., Seavey, K.C., 2006. Sigma-Profile Database for Using COSMO-Based Thermodynamic Methods. https://doi.org/10.1021/ie060370h.
Peng, D., Zhang, J., Cheng, H., Chen, L., Qi, Z., 2017. Computer-aided ionic liquid design for separation processes based on group contribution method and COSMO-SAC model. Chem. Eng. Sci. 159, 58–68. https://doi.org/10.1016/j.ces.2016.05.027.
Peric, B., Sierra, J., Martí, E., Cruañas, R., Garau, M.A., 2015. Quantitative structure-activity relationship (QSAR) prediction of (eco)toxicity of short aliphatic protic ionic liquids. Ecotoxicol. Environ. Saf. 115, 257–262. https://doi.org/10.1016/j.ecoenv.2015.02.027.
Ranke, J., Mölter, K., Stock, F., Bottin-Weber, U., Poczobutt, J., Hoffmann, J., Ondruschka, B., Filser, J., Jastorff, B., 2004. Biological effects of imidazolium ionic liquids with varying chain lengths in acute Vibrio fischeri and WST-1 cell viability assays. Ecotoxicol. Environ. Saf. 58, 396–404. https://doi.org/10.1016/S0147-6513(03)00105-2.
Ranke, J., Müller, A., Bottin-Weber, U., Stock, F., Stolte, S., Arning, J., Störmann, R., Jastorff, B., 2007. Lipophilicity parameters for ionic liquid cations and their correlation to in vitro cytotoxicity. Ecotoxicol. Environ. Saf. 67, 430–438. https://doi.org/10.1016/j.ecoenv.2006.08.008.
Roy, K., Kar, S., Ambure, P., 2015. On a simple approach for determining applicability domain of QSAR models. Chemom. Intell. Lab. Syst. 145, 22–29. https://doi.org/10.1016/J.CHEMOLAB.2015.04.013.
Rybka, M., Mercader, A.G., Castro, E.A., 2014. Predictive QSAR study of chalcone derivatives cytotoxicity activity against HT-29 human colon adenocarcinoma cell lines. Chemom. Intell. Lab. Syst. 132, 18–29. https://doi.org/10.1016/j.chemolab.2013.12.005.
Sanchez Zayas, M., Gaitor, J.C., Nestor, S.T., Minkowicz, S., Sheng, Y., Mirjafari, A., 2016. Bifunctional hydrophobic ionic liquids: facile synthesis by thiol–ene “click” chemistry. Green Chem. 18, 2443–2452. https://doi.org/10.1039/C5GC01930C.
Simon, L., Abdelmalek, B., 2012. Design of Skin Penetration Enhancers Using Replacement Methods for the Selection of the Molecular Descriptors. Pharmaceutics 4, 343–353. https://doi.org/10.3390/pharmaceutics4030343.
Singh, K.P., Gupta, S., Basant, N., 2014. Predicting toxicities of ionic liquids in multiple test species-an aid in designing green chemicals †. RSC Adv. https://doi.org/10.1039/c4ra11252k.
Smirnova, N.A., Safonova, E.A., 2010. Ionic liquids as surfactants. Russ. J. Phys. Chem. A 84, 1695–1704. https://doi.org/10.1134/S0036024410100067.
Song, Z., Zhang, J., Zeng, Q., Cheng, H., Chen, L., Qi, Z., 2016. Effect of cation alkyl chain length on liquid-liquid equilibria of {ionic liquids + thiophene + heptane}: COSMO-RS prediction and experimental verification. Fluid Phase Equilib. 425, 244–251. https://doi.org/10.1016/j.fluid.2016.06.016.
Sosnowska, A., Grzonkowska, M., Puzyn, T., 2017. Global versus local QSAR models for predicting ionic liquids toxicity against IPC-81 leukemia rat cell line: The predictive ability. J. Mol. Liq. 231, 333–340. https://doi.org/10.1016/j.molliq.2017.02.025.
Sun, M., Zheng, Y., Wei, H., Chen, J., Cai, J., Jin, M., 2009. Enhanced replacement method-based quantitative structure-activity relationship modeling and support vector machine classification of 4-anilino-3-quinolinecarbonitriles as Src kinase inhibitors. QSAR Comb. Sci. 28, 312–324. https://doi.org/10.1002/qsar.200860107.
The UFT/ Merck Ionic Liquids Biological Effects Database, http://www.il-eco.uft.uni-bremen.de.
Torrecilla, J.S., García, J., Rojo, E., Rodríguez, F., 2009. Estimation of toxicity of ionic liquids in Leukemia Rat Cell Line and Acetylcholinesterase enzyme by principal component analysis, neural networks and multiple lineal regressions. J. Hazard. Mater 164, 182–194. https://doi.org/10.1016/J.JHAZMAT.2008.08.022.
Torrecilla, J.S., Palomar, J., Lemus, J., Rodríguez, F., 2010. A quantum-chemical-based guide to analyze/quantify the cytotoxicity of ionic liquids †. Green Chem. 12, 123–134. https://doi.org/10.1039/b919806g.
Tot, A., Vraneš, M., Maksimović, I., Putnik-Delić, M., Daničić, M., Belić, S., Gadžurić, S., 2018. The effect of imidazolium based ionic liquids on wheat and barley germination and growth: Influence of length and oxygen functionalization of alkyl side chain. Ecotoxicol. Environ. Saf. 147, 401–406. https://doi.org/10.1016/j.ecoenv.2017.08.066.
Ventura, S.P.M., Gonçalves, A.M.M., Sintra, T., Pereira, J.L., Gonçalves, F., Coutinho, J.A.P., 2013. Designing ionic liquids: the chemical structure role in the toxicity. Ecotoxicology 22, 1–12. https://doi.org/10.1007/s10646-012-0997-x.
Wlazło, M., Karpińska, M., Domańska, U., 2017. Separation of water/butan-1-ol mixtures based on limiting activity coefficients with phosphonium-based ionic liquid. J. Chem. Thermodyn. 113, 183–191. https://doi.org/10.1016/j.jct.2017.06.011.
Yan, F., Shang, Q., Xia, S., Wang, Q., Ma, P., 2015. Topological study on the toxicity of ionic liquids on Vibrio fischeri by the quantitative structure–activity relationship method. J. Hazard. Mater 286, 410–415. https://doi.org/10.1016/J.JHAZMAT.2015.01.016.
Yan, F., Xia, S., Wang, Q., Ma, P., 2012. Predicting the Decomposition Temperature of Ionic Liquids by the Quantitative Structure–Property Relationship Method Using a New Topological Index. J. Chem. Eng. Data 57, 805–810. https://doi.org/10.1021/je201023a.
Zhang, S., Sun, N., He, X., Lu, X., Zhang, X., 2006. Physical Properties of Ionic Liquids: Database and Evaluation. J. Phys. Chem. Ref. Data 35, 1475–1517. https://doi.org/10.1063/1.2204959.
Zhao, Y., Zeng, S., Huang, Y., Afzal, R.M., Zhang, X., 2015. Estimation of Heat Capacity of Ionic Liquids Using S σ-profile Molecular Descriptors. Ind. Eng. Chem. Res. 54, 12987–12992. https://doi.org/10.1021/acs.iecr.5b03576.
Zhao, Y., Zhao, J., Huang, Y., Zhou, Q., Zhang, X., Zhang, S., 2014. Toxicity of ionic liquids: Database and prediction via quantitative structure-activity relationship method. J. Hazard. Mater 278, 320–329. https://doi.org/10.1016/j.jhazmat.2014.06.018.
Zhou, T., Wang, Z., Ye, Y., Chen, L., Xu, J., Qi, Z., 2012. Deep separation of benzene from cyclohexane by liquid extraction using ionic liquids as the solvent. Ind. Eng. Chem. Res. 51, 5559–5564. https://doi.org/10.1021/ie202728j.